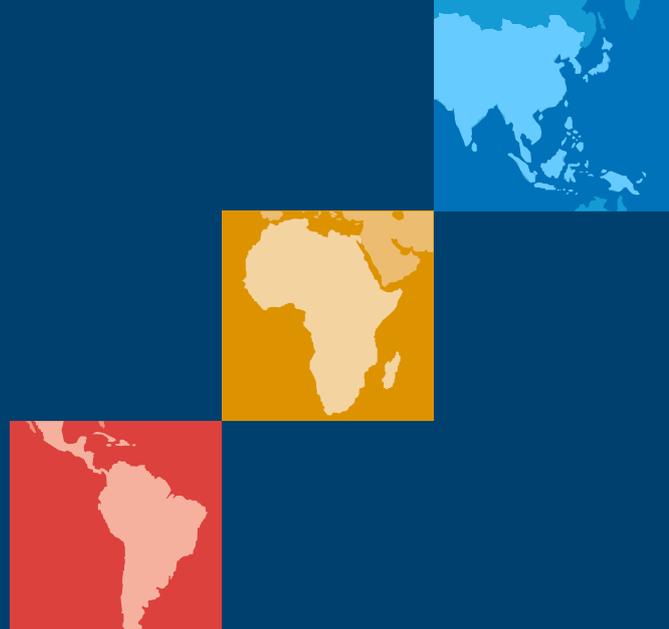


DISCUSSION PAPER / 2010.01



Challenges in impact evaluation of development interventions: opportunities and limitations for randomized experiments

Jos Vaessen



University
of Antwerp



INSTITUTE OF DEVELOPMENT
POLICY AND MANAGEMENT

Comments on this Discussion Paper are invited.
Please contact the author at: jos.vaessen@ua.ac.be

Instituut voor Ontwikkelingsbeleid en -Beheer
Institute of Development Policy and Management
Institut de Politique et de Gestion du Développement
Instituto de Política y Gestión del Desarrollo

Postal address:
Prinsstraat 13
B-2000 Antwerpen
Belgium

Visiting address:
Lange Sint-Annastraat 7
B-2000 Antwerpen
Belgium

Tel: +32 (0)3 275 57 70
Fax: +32 (0)3 275 57 71
e-mail: dev@ua.ac.be
<http://www.ua.ac.be/dev>

DISCUSSION PAPER / 2010.01

Challenges in impact evaluation of development interventions: opportunities and limitations for randomized experiments

Jos Vaessen*

August 2010

* Jos Vaessen is lecturer at Maastricht University and researcher at the Institute of Development Policy and Management (IOB), University of Antwerp.

TABLE OF CONTENTS

	ABSTRACT	6
	RÉSUMÉ	7
1.	THE RISING STAR OF IMPACT EVALUATION IN DEVELOPMENT	8
2.	SCOPE	11
3.	CHALLENGE 1: DELIMITATION	13
3.1.	The importance of stakeholder values	13
3.2.	The impact of what?	14
3.3.	The impact on what?	16
3.3.1.	Institutional versus beneficiary level effects	16
3.3.2.	Intended versus unintended effects	17
3.3.3.	Short-term versus long-term effects	18
4.	CHALLENGE 2: ATTRIBUTION ‘VERSUS’ EXPLANATION	19
4.1.	Addressing the attribution challenge with randomized experiments	19
4.2.	Internal versus external validity	20
4.3.	Interventions as theories	22
4.4.	The law of comparative advantages: theory-based and multi-method evaluation	23
5.	CHALLENGE 3: IMPACT EVALUATION IN PRACTICE	25
5.1.	Threats to attribution analysis in experimental settings	25
5.2.	Other design and implementation challenges	26
6.	SOME LESSONS FOR RANDOMIZED EXPERIMENTS AND IMPACT EVALUATION IN GENERAL FROM THE PERSPECTIVE OF A ‘NON-RANDOMISTA’	29
	REFERENCES	34

ABSTRACT

In recent years debates on as well as funding of impact evaluations of development interventions have flourished. Unfortunately, controversy regarding the promotion and application of randomized experiments (RE) has led to a sense of polarization in the development policy and evaluation community. As some proponents claim epistemological supremacy of REs (with respect to attribution) the counter reaction among others has been rejection. Needless to say, such extreme positions are counterproductive to reaching a goal that is commonly endorsed: to learn more about what works and why in development. This paper discusses the prospects and limitations of REs from the perspective of three categories of challenges in impact evaluation: delimitation and scope, attribution versus explanation, and implementation challenges. The implicit lesson is twofold. First of all, the question ‘to randomize or not to randomize’ is overrated in the current debate. Limitations in scope, applicability as well as implementation will necessarily restrict the use of REs in development impact evaluation. There is a risk that the current popularity of REs in certain research and policy circles might lead to a backlash as too high expectations of REs may quicken its demise. More importantly, given the nature and scope of the challenges discussed in the paper, more energy should be devoted to developing and testing ‘rigorous’ mixed method approaches within a framework of theory-driven evaluation.

Acknowledgments

I would like to thank Frans Leeuw and Robrecht Renard for their comments on this paper. Any remaining errors are my own.

RÉSUMÉ

Ces dernières années, les évaluations d'impact des interventions en matière de développement ont fait l'objet de nombreux débats et ont été largement financées. Malheureusement, la controverse au sujet de la promotion et de l'application des expériences randomisées (ER) a suscité un sentiment de polarisation parmi ceux qui définissent et évaluent les politiques de développement. Comme certains partisans revendiquent la suprématie épistémologique des ER (pour ce qui est de l'attribution), la réaction de certains autres a pris la forme d'un rejet. Il va sans dire que ces positions extrêmes sont contreproductives dans la poursuite d'un objectif partagé par tous : en savoir plus long sur ce qui fonctionne en matière de développement, et pourquoi. Ce document commente les perspectives et les limitations des ER du point de vue de trois catégories de défis dans l'évaluation d'impact : délimitation et étendue, attribution versus explication, et les défis liés à la mise en œuvre. La leçon implicite est double. Premièrement, la question 'randomiser ou non' prend trop d'importance dans le débat actuel. Des limitations d'étendue, d'applicabilité ainsi que de mise en œuvre restreindront fatalement l'utilisation des ER dans l'évaluation de l'impact du développement. Il se pourrait que la popularité actuelle des ER dans certains cercles de chercheurs et de décideurs entraîne un retour de bâton, car des attentes trop élevées par rapport aux RE pourraient accélérer leur abandon. Enfin, et surtout, étant donné la nature et l'étendue des défis commentés dans ce document, il est nécessaire de consacrer plus d'énergie à l'élaboration et à la mise à l'épreuve d'approches mixtes 'rigoureuses' dans un cadre d'évaluation basé sur la théorie.

1. THE RISING STAR OF IMPACT EVALUATION IN DEVELOPMENT

The question of ‘what works and why’ in development assistance has received considerable attention over the past few years. The major reason is that many outside of development agencies believe that achievement of results has been poor, or at best not convincingly established. In the last decades of the previous century part of the development assistance paradigm was about ‘thinking big’, e.g. structural adjustment policies as key factors in generating stability and growth in developing countries. Correspondingly, a lot of intellectual effort went into analyses of development at the macro level, with scores of economists working on growth regressions, trying to identify the key factors that were determining country growth of GDP and the role of development assistance therein. Growing pessimism within the international community about the effectiveness of macro interventions, in part fuelled by the lack of decisive evidence from the academic community, gradually led to a shift in development paradigm towards more ‘thinking small’ (Easterly, 2001; Cohen and Easterly, 2009). Yet, the state of evidence on the effectiveness of concrete policy interventions (programs, policies) was far from promising either. Towards the end of the previous century the realization grew that, given the evidence base, it was hard to determine the extent to which interventions were making a difference (Baker, 2000). In 2006, an influential paper published by the Center for Global Development –“When will we ever learn?” (CGD, 2006)- pointed at an evaluation gap in development; despite enormous investments in development policy, the evidence base on what works was diagnosed as weak. According to the paper, too much of the evaluative work in development focused on process instead of results and credible evaluations of results were scarce. Fortunately, at the time of publication of this paper, the tide had already been gradually turning. A number of key events such as the endorsement of the Millennium Development Goals by the global community in 2001, the 2002 Monterrey Conference on Financing for Development, the 2005 Paris Declaration on Development Effectiveness and the 2008 High-Level Forum on Aid Effectiveness in Accra were signs of a growing results-focus in the development community, gradually paving the road for more attention to the assessment of effects of development interventions.

As a result of this evolution, in recent years debates on as well as funding of impact evaluation have flourished. Impact evaluation can be roughly defined as the (growing) field of evaluative practices aimed at assessing the intended and unintended effects of policy interventions. One of the particularly productive areas in impact evaluation is (quasi-)experimental impact evaluation, in particular randomized experiments. A number of initiatives, most notably the Poverty Action Lab (J-PAL), Innovations for Poverty Action and the World Bank’s Spanish Impact Evaluation Fund, are creating a growing body of evaluative evidence based on randomized experiments.[1] The comparative advantage of the latter methodology,[2] as it has been argued widely, is its inherent strength to address the attribution problem in evaluation through counterfactual analysis. The basic idea of a randomized experiment (RE)[3] is that the situation of a participant group (receiving benefits from/affected by an intervention) is com-

[1] Another fairly recent initiative, the International Initiative for Impact Evaluation, is funding proposals for research based on a broader palette of methodological designs, usually a combination of quantitative methods embedded in a theory-based approach.

[2] To keep things simple, I distinguish between methodologies and methodological designs on the one hand and methods on the other. The former can comprise multiple methods, the latter referring to specific tools of data collection and analysis.

[3] Given the scope of the interventions that are currently assessed with this methodology, I prefer the term randomized experiments (RE) instead of the “treatment-oriented” term of randomized controlled trials.

pared over time with the situation of an equivalent control group that is not affected by the intervention. Allocation to either of these groups^[1] is random. Consequently, in sufficiently large samples the probability that both groups are equivalent on all observable and non-observable characteristics except for intervention participation is very high (see for example Shadish et al., 2002; Morgan and Winship, 2007). This inherent strength of REs can resolve the selection bias problem in evaluation. People that participate or are affected by an intervention usually differ from the population at large due to self-selection or targeting. As a result, simple comparisons between participants and people not covered by the intervention will be biased. Randomization of intervention benefits or participation addresses this issue. This particular feature of REs has been lauded by growing groups of researchers and decisionmakers in development. Indeed, for some the rise of REs signifies the beginning of a new era in development evaluation: “[c]reating a culture in which rigorous randomized evaluations are promoted, encouraged, and financed has the potential to revolutionize social policy during the 21st century, just as randomized trials revolutionized medicine during the 20th” (The Lancet Editorial quoting Esther Duflo, 2004: 731). Recent examples of randomized experiments include Miguel and Kremer (2004) on deworming treatment in Kenya, Banerjee et al. (2007) on education in India, Olken (2007) on monitoring corruption in Indonesia, McKenzie and Woodruff (2008) on returns to capital and access to finance in Mexico, or Karlan and Zinman (2009) on microcredit in the Philippines.

While the growing body of evidence based on REs certainly seems promising, the fact that some of the protagonists of the RE tradition, dubbed ‘randomistas’, perceive REs as the only way to produce rigorous evaluative evidence has led to a storm of critique.^[2] Ravallion (2009a: 1) warns about the consequences of the growing influence of the ‘randomistas’: “Researchers are turning down opportunities to evaluate public programs when randomization is not feasible. Doctoral students are searching for something to randomize. Philanthropic agencies are sometimes unwilling to fund non-experimental evaluations.” A key argument against the alleged supremacy of REs^[3] expressed by critics (see below), has been that if one randomly controls for all observable and non-observable confounders, one cannot generalize conclusions about effectiveness beyond the specific sample, as one does not know exactly in what aspects the experimental sample differs from the population at large (see for example Deaton, 2009).

The phenomenon of REs being perceived as superior to other methods of impact evaluation by a part of the development community is not new and has also occurred in other disciplines and policy fields. REs have been applied since the early twentieth century in the fields of education, crime, health and social welfare mostly by psychologists, sociologists, economists and health scientists (see Oakley, 2000; Leeuw, 2009). Especially in evidence-based medicine REs have gained a prominent role, yet have also been criticized (see for example Worrall, 2007). In the field of international development, REs have been embraced mostly by economists. Rodrik (2008) and Deaton (2009) argue that this is due to mainly two reasons. First, they note a certain sense of disappointment among economists regarding the failure of economic theory in providing accurate and generalizable guidance on effectiveness. Second, there is growing consensus on

[1] In fact, there may be more than two groups, as multiple “treatment” groups may be defined and monitored over time.

[2] See Cohen and Easterly (2009) for a collection of essays by eminent development scholars in favor and against REs as principal tools for assessing development effectiveness.

[3] Including the corresponding hierarchy of methodological designs that places randomized controlled trials at the top of the “food chain” followed by several quasi-experimental designs. A well-known example of such a hierarchy of designs is the Maryland Scale of Scientific Methods (Sherman et al., 1997).

the shortcomings in econometric methods for explaining effectiveness, mainly due to the identification problem.^[1] As REs have been enthusiastically taken up by development economists, sharp critique has come from the same discipline, most notably from Rodrik (2008), Ravallion (2008, 2009a) and Deaton (2009).^[2] Banerjee and Duflo (2008), two prominent ‘randomistas’, acknowledge much of the critique on REs. However, at the same time they counter the critique by arguing that most of it is not specific to randomized experiments but also applicable to non-experimental observational studies. However, this is exactly a point that critics refer to. The fact that randomized experiments can be justifiably criticized on a number of key issues (see below) undermines the alleged claim to epistemological supremacy, a claim that many ‘randomistas’ explicitly or implicitly endorse.

[1] Generally, this refers to the problem that multiple values of parameters or multiple models might explain a certain pattern in the data.

[2] See also Cohen and Easterly (2009).

2. SCOPE

The controversy surrounding REs unfortunately has kindled a sense of polarization within the development research and evaluation community, and indeed also within practitioner and policymaker circles (see for example Cohen and Easterly, 2009), which is counterproductive to achieving the important goal that so many endorse: to promote a growth of knowledge on what works and why in development. In this paper, when discussing the multiple conceptual and methodological challenges in impact evaluation, I will recurrently refer to REs and their potential to address a particular challenge. The discussion will demonstrate that REs are not equipped to address all of these challenges. However, this should not lead to erroneous conclusions about the utility of REs. Whereas the potential role of REs (and quasi-experimental designs)^[1] in impact evaluation is probably overestimated by ‘randomistas’ and in some cases also policymakers, they do possess a comparative advantage in addressing the issue of attribution (internal validity).

As a structure to the discussion I discern three key challenges in impact evaluation, each of which is further classified into sub issues. The classification is based on three pillars. First of all, it relies on an elaborate discussion in 2008-2009 with a community of practitioners and scholars working on impact evaluation in development.^[2] The second pillar concerns a review of the current literature on impact evaluation and development. Finally, my own empirical and conceptual work on impact evaluation also served as inspiration for the structure employed below.

Challenge 1: Delimitation. The scope of an impact evaluation can widely differ depending on the nature of the evaluand, the types of effects that might occur, as well as the choices that are made about the aspects to be assessed in detail. These choices can be determined by decisionmakers and/or researchers and may include the priorities of other stakeholder groups such as target groups.

Challenge 2: Attribution versus explanation. REs have a comparative advantage in determining the net effects^[3] of an intervention with a high degree of internal validity. Theory can help to strengthen the internal validity of findings by elucidating how and why certain changes occur. In addition, theory can strengthen the external validity of findings.^[4]

[1] Quasi-experiments do not rely on randomization but on other principles for establishing participant and control groups (which are generally considered as inferior to randomization in terms of their potential to generate unbiased estimates of impact).

[2] In 2008-2009, the Network of Networks for Impact Evaluation (<http://www.worldbank.org/ieg/nonie>, last consulted January 11, 2009) commissioned an assignment with the purpose of revising and adding new content to existing guidelines on impact evaluation. The result was a new Guidance on impact evaluation (Leeuw and Vaessen, 2009).

[3] This refers to the effects of an intervention adjusted for what would have happened if the intervention had not taken place, a concept often used in counterfactual analysis.

[4] Internal validity is about establishing a causal relationship between intervention outputs and processes of change leading to outcomes and impacts. External validity concerns the generalizability of findings to other settings. In addition to these two dimensions it can also be argued that theory can strengthen the construct validity (i.e. the extent to which variables adequately represent the phenomena they refer to) and statistical conclusion validity (i.e. the degree of confidence about the existence of a relationship between intervention and effect variable) of findings (see for example Cook and Campbell, 1979; Shadish et al., 2002).

Challenge 3: Impact evaluation in practice. Good impact evaluation is good research. Whereas many of the current debates on impact evaluation tend to center on arguments pro and against REs, in practice the validity of findings of any type of impact evaluation, including REs, heavily depends on the extent to which a number of key design and implementation challenges have been appropriately addressed.

In this paper the focus will be on methodological design and implementation aspects of impact evaluation. Less attention will be paid to the properties of statistical analysis within the context of REs or other methodological designs (see for example Heckman, 1992; Shadish et al. 2002; Cook, 2006; for development interventions see for example Deaton, 2009).

3. CHALLENGE 1: DELIMITATION

3.1. The importance of stakeholder values

A first important criterion for delimitation in impact evaluation concerns the question of ‘impact according to whom’. There is a strong movement in development research and practice which endorses the idea that impact evaluation is not only about assessing the effects of an intervention but also about underlying questions of what types of processes of change and effects are valued as important (either positive or negative) and by whom?^[1]

This line of thought is most manifest in the evaluation tradition of participatory impact evaluation. Nowadays, participatory methods have become ‘mainstream’ tools in development in almost every area of policy intervention. The roots of participation in development lie in the rural sector, where Chambers (1995) and others developed the now widely used principles of Participatory Rural Appraisal.^[2] Participatory evaluation approaches (see for example, Cousins and Whitmore, 1998) are built on the principle that stakeholders should be involved in some or all stages of the evaluation. In the case of impact evaluation participation includes aspects such as the determination of objectives, indicators to be taken into account, as well as stakeholder participation in data collection and analysis. In practice it can be useful to differentiate between stakeholder participation as a process and stakeholder perceptions and views as sources of evidence (Cousins and Whitmore, 1998).

Randomized designs are not about stakeholder participation or elicitation of stakeholder values. However, there is no reason to assume that REs may not be combined with participatory processes and methods of data collection (Karlan, 2009). In practice a wide variety of methods are available (see for example IFAD, 2002; Mikkelsen, 2005; Pretty et al., 1995; Salmen and Kane, 2006). Stakeholder participation in impact evaluation can be beneficial in many ways, i.e. by enhancing the ownership and (possibly) utilization of an evaluation, improving the quality of the data collected from target populations, or strengthening local processes of governance. At the same time, participatory methods have been criticized on many grounds. Often mentioned critical aspects concern the limited applicability of impact evaluations with a high degree of participation especially in large-scale, comprehensive, multi-site interventions. In such contexts, organizing processes of stakeholder participation may not be feasible due to high costs and logistical barriers. In addition, there is some doubt about the reliability of information based on stakeholder perceptions (e.g. due to risks of strategic responses, manipulation of information or advocacy by stakeholders).^[3]

[1] The importance of stakeholder values may differ according to the type of intervention. For example, whereas stakeholder views on what is important in the evaluation of the effects of reforestation programs may widely differ, this may be less the case for health indicators in the case of nutrition programs.

[2] Participatory Impact Assessment is an extension of Participatory Rural Appraisal and involves the adaptation of participatory tools combined with more conventional statistical approaches specifically to measure the impact of humanitarian assistance and development projects on people’s lives (Catley et al., 2008).

[3] In turn, proponents of participatory evaluation approaches are often very critical of the reliability of survey-based research and corresponding data analyses within the framework of experimental or non-experimental methodological designs. Especially well-known is Robert Chambers’ (1983) discussion of what he labeled as ‘survey slavery’, criticizing among other things the costs (also for respondents), inefficiencies, rigidities and data problems of survey research in rural development contexts.

An alternative approach for eliciting stakeholder values which does not rely on stakeholder participation is values inquiry and “refers to a variety of methods that can be applied to the systematic assessment of the value positions surrounding the existence, activities, and outcomes of a social policy and program” (Mark et al., 1999: 183). Values inquiry exercises may be more useful than participatory evaluation approaches in situations where policy makers are interested in a representative picture of the value positions of large groups of beneficiaries dispersed over large territories.

3.2. The impact of what?

When talking about the scope and delimitation of impact evaluation it is useful to address the following two questions: the impact of what and the impact on what (see also Vaessen and Todd, 2008)? Regarding the impact of what, today more than ever one can speak of a ‘continuum’ of interventions. At one end of the continuum are relatively simple projects characterized by ‘single strand’ initiatives with explicit objectives, carried out within a relatively short timeframe, where interventions can be isolated, manipulated and measured. An impact evaluation in the agricultural sector for example, will seek to attribute changes in crop yield to an intervention such as a new technology or agricultural practice. In a similar guise, in the health sector, a reduction in malaria will be analyzed in relation to the introduction of bed nets. For these types of interventions, experimental and quasi-experimental designs may be appropriate for assessing causal relationships. At the other end of the continuum are comprehensive programs with an extensive range and scope (increasingly at country, regional or global level), with a variety of activities that cut across sectors, themes and geographic areas, and emergent specific activities. Many of these interventions address aspects that are assumed to be critical for effective development yet difficult to define and measure, such as human security, good governance, political will and capacity, sustainability, and effective institutional systems.

One of the trends in development is that donors are moving up the ‘aid chain’. Whereas in the past donors were very much involved in ‘micro-managing’ their own projects and (sometimes) bypassing government systems, nowadays a sizeable chunk of aid is allocated to national support for recipient governments. Attention to some extent has shifted from micro-earmarking (e.g. donor money destined for an irrigation project in district x) to meso-earmarking (e.g. support for the agricultural sector) or macro-earmarking (e.g. support for the government budget to be allocated according to country priorities). Besides a continued interest in the impact of individual projects, donors, governments and nongovernmental institutions are increasingly interested in the impact of comprehensive programs, sector strategies or country strategies, often comprising multiple instruments, stakeholders, sites of intervention and target groups (see Jones et al. (2008) for a recent inventory of impact evaluations in different sectors of development intervention).

In most countries donor organizations are (still) the main promoters of impact evaluation. The partial shift of the unit of analysis to the macro and (government) institutional level requires impact evaluators to pay more attention to complicated and more complex interventions at national, sector or program level. Multi-site, multi-governance and multiple (simultaneous) causal strands are important elements of this (see Rogers, 2008). At the same time, the need for more rigorous impact evaluation at the level of ‘single strand’ projects or activities remains as important as ever since they are the building blocks of higher-level programs and

policies. Furthermore, the ongoing efforts in capacity-building on national M&E systems (see Kusek and Rist, 2004; Morra and Rist, 2009) and the promotion of country-led evaluation efforts stress the need for further guidance on impact evaluation at ‘single intervention’ level.

Within the light of the heterogeneous landscape of interventions, critics and (most) proponents of REs alike acknowledge the limitations in applicability of REs. What is problematic is that the special status attributed to REs by some researchers and policymakers is likely to generate a bias in terms of too much evaluative focus on interventions that are amenable to this approach. Already development evaluation is biased in terms of what Ravallion calls a “‘myopia bias’ in our knowledge, [with evaluation] favoring development projects that yield quick results” (Ravallion, 2008: 6). Similarly, Blattman (2008) refers to the ‘overevaluation’ of certain economic, educational and health interventions and the ‘underevaluation’ of interventions on peace-building, crime reduction, and governance issues (e.g. public management, decentralization; see also Jones et al., 2008). REs are most readily applicable in case of discrete, homogenous interventions with clearly delineated target groups^[1] rather than more complicated interventions, interventions that evolve during implementation or full-coverage interventions such as laws or macroeconomic policies (Bamberger and White, 2007; Rossi et al. 2004). Even in the case of rather simple interventions, quantitative researchers such as those handling REs find themselves opposed by anthropologists and sociologists who criticize the rather simplistic view of deconstructing interventions into neatly delineated packages of activities, benefits or treatments (see for example Hulme (2000) for a discussion). In contrast, they emphasize the complexity in social development interventions. Staff in such interventions “are not functionaries dutifully providing a standardised service, such as immunising babies or distributing food rations; they are instead engaged in extensive face-to-face interaction with villagers over many months, making innumerable discretionary decisions. In many respects ‘the project’ is itself a dynamic decision-making process rather than a static ‘product’, and as such attempts to make causal claims regarding overall impact must address endemic unobserved heterogeneity bias. In short, on both the ‘demand side’ (local context) and the ‘supply side’ (front-line project staff) there is, by design, enormous variation” (Woolcock, 2009: 5). Indeed, identifying what the intervention exactly is, where it begins and where it ends can be rather challenging (Pawson and Tilley, 1997).

In case of complicated interventions (comprising multiple (interacting) intervention components) at best sometimes only some of the components may be amenable to a RE. In case of interventions that focus on the institutional level (e.g. capacity-building, technical assistance, administrative reform), corresponding evaluations look at one or a few units of analysis (i.e. the institution) –Bamberger and White (2007) call this the small n problem- a situation that is not amenable to a RE. Homogeneity of the intervention is another frequently referred to condition for REs. A proper RE requires a high degree of homogeneity in intervention, target groups and context over time, conditions which are unlikely to hold in many cases. Often the nature of the intervention changes over time through adaptive learning, political pressures or to obtain more funding (e.g. adding training to subsidy programs). In addition, there might be changes in contextual factors that affect participants differently than control group members. For example, rising output prices might speed up technology adoption processes among participants of a training program more than in the case of control group members who are facing a knowledge

[1] This in line with what Ravallion (2009b) calls assigned programs.

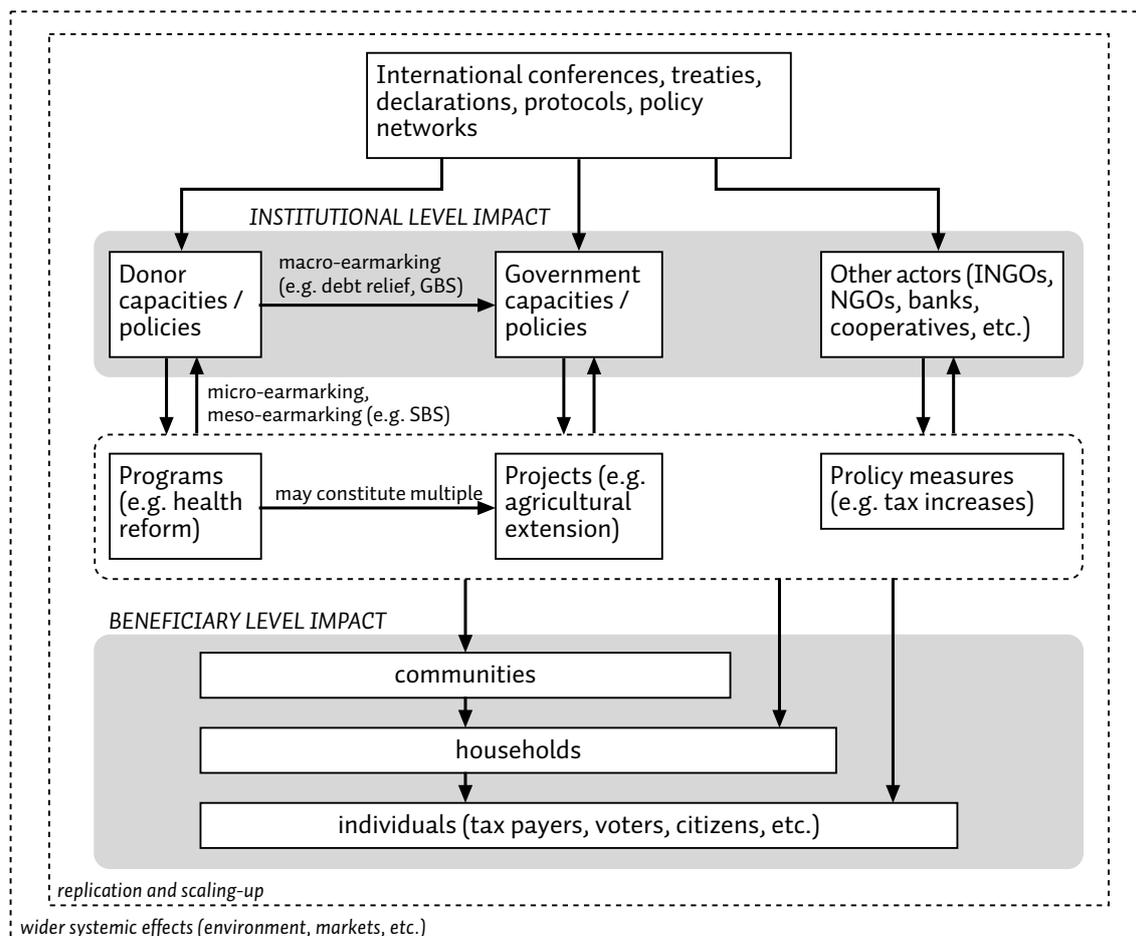
constraint. Given the above, if anything, a narrow focus on REs would reinforce an already existing evaluation bias towards particular types of interventions.

3.3. The impact on what?

3.3.1. Institutional versus beneficiary level effects

The second issue, the impact on what, concerns the type of effects that we are interested in. The causality chain linking policy interventions to ultimate policy goals (e.g. poverty alleviation) can be relatively direct and straightforward (e.g. the impact of vaccination programs on mortality levels) but also complex and diffuse. Impact evaluations of for example sector strategies or general budget support potentially encompass multiple causal pathways resulting in long-term direct and indirect impacts. Some of the causal pathways linking interventions to impacts may be fairly straightforward (e.g. from training programs in alternative income generating activities to employment and to income levels), whereas other pathways are more complex and diffuse in terms of going through more intermediate changes, and being contingent upon more external variables (e.g. from stakeholder dialogue to changes in policy priorities to changes in policy implementation to changes in human welfare).

Figure 1. Levels of intervention, programs and policies and types of impact



Source: Leeuw and Vaessen (2009)

Given this diversity it is useful for purposes of ‘scoping’ to distinguish between two principal levels of impact: impact at the institutional level and impact at the beneficiary level (see Figure 1).^[1] It broadens impact evaluation beyond either measuring whether objectives have been achieved or assessing direct effects on intended beneficiaries. It includes the full range of impacts at all levels of the results chain, including ripple effects on families, households and communities, on institutional, technical or social systems, and on the environment. Interventions that can be labeled as institutional primarily aim at changing second-order conditions (i.e. the capacities, willingness, and organizational structures enabling institutions to design, manage and implement better policies for communities, households and individuals). Examples are policy dialogues, policy networks, training programs, institutional reforms, and strategic support to institutional actors (i.e. governmental, civil society institutions, private corporations, hybrids) and public private partnerships. Other types of interventions directly aim at/affect communities, households, individuals, including voters and taxpayers. Examples are fiscal reforms, trade liberalization measures, technical assistance programs, cash transfer programs, construction of schools, etc.

3.3.2. Intended versus unintended effects

A widely endorsed reference to impact evaluation concerns the OECD-DAC definition, which defines impacts as (OECD-DAC, 2002: 24) “[p]ositive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended”. However, when we look at the body of research under the banner of impact evaluation, a large part of it is not on long-term results, nor on indirect and unintended results. In fact, the body of evaluation research based on REs and quasi-experiments is usually about analyzing the attribution of short-term outcomes to a particular intervention (White, 2009). Moreover, in order to permit a difference in difference estimation of the net effects of an intervention, only indicators of effects that are expected to be important are taken into account. As a result of the rather limited and rigid set of indicators employed within the framework of REs and quasi-experiments they are not very useful for identifying and analyzing unanticipated effects (see for example Davidson, 2006; Bamberger and White, 2007).

Policymakers are very much interested in the indirect diffusion, replication or scaling-up effects of interventions. Whether intended or unintended they usually are ‘under’ the radar of REs as the analysis of these effects requires broadening the scope of indicators as well as the sampling framework of an impact evaluation. Replicatory effects in terms of behavioral changes in actors beyond the original target group can stem from market responses (given that participants and non-participants trade in the same markets), the (non-market) behavior of participants/non-participants or the behavior of intervening agents (governmental/NGO). For example, “aid projects often target local areas, assuming that the local government will not respond; yet if one village gets the project, the local government may well cut its spending on that village, and move to the control village” (Ravallion, 2008: 10). Another example concerns displacement effects of environmentally damaging land use towards other areas beyond the grasp of an intervention; deforestation may increase elsewhere as land use becomes more restricted in certain areas.

[1] In addition, one can discern other “levels” such as replicatory effects and systemic effects.

3.3.3. Short-term versus long-term effects

In some types of interventions, effects emerge quickly. In others effects may take much longer to become manifest, and change over time. The timing of the evaluation is therefore important. Development interventions are usually assumed to contribute to long-term development (with the exception of humanitarian disaster and emergency situations). However, focusing on short-term or intermediate outcomes often provides for more useful and immediate information for policy- and decision-making. However, intermediate outcomes may be misleading, often differing markedly from those achieved in the longer term. Many of the impacts of interest from development interventions will only be evident in the longer-term, such as environmental changes, or effects on subsequent generations. Searching for evidence of such effects too early might mistakenly conclude that interventions have failed.

In this context, the exposure time of an intervention to be able to make an impact is an important point. A typical agricultural innovation project that tries to change farmers' behavior with incentives (training, technical assistance, credit) is faced with time lags in both the adoption effect (farmers typically are risk averse and face resource constraints and start adopting innovations on an experimental scale) as well as the diffusion effect (other farmers want to see evidence of results before they copy). In such gradual non-linear processes of change with cascading effects, the timing of the ex post measurement (of land use) is crucial. Ex post measurements just after project closure could either underestimate (full adoption/diffusion of interesting practices has not taken place yet) or overestimate impact (as farmers will stop investing in those land use practices that are not attractive enough to be maintained without project incentives).

Woolcock (2009) has recently highlighted a related problem. He argues that REs, and especially those that are based on a limited number of data points in time (e.g. before and after only), do not take into account the nature and dynamics of processes of change induced by an intervention. Processes of change are often not linear. In practice, processes of change often resemble j-curves or step functions. Examples of such processes are the effects of microcredit on empowerment (e.g. initial resistance by men until persistent and collective pressures lead to a shift in norms) or the adoption of new agricultural technologies (e.g. Rogers, 2003). The implication for REs is that for example if ex post measurement happens to take place when the change curve has hit a (temporal) low, then estimates of net effects will be entirely unrealistic. REs that are not supported by theory or data from additional methods of inquiry are not equipped to address the abovementioned issues.^[1]

[1] It is important to mention that REs using a simple difference in difference estimate of net impact (ex ante versus ex post) are more prone to error than designs that rely on multiple waves of measurement within participant and control groups.

4. CHALLENGE 2: ATTRIBUTION ‘VERSUS’ EXPLANATION

4.1. Addressing the attribution challenge with randomized experiments

The OECD-DAC definition of impacts mentioned above refers to the ‘effects produced by’ an intervention, stressing the attribution aspect. This implies an approach to impact evaluation which is about attributing impacts rather than ‘merely’ assessing what happened.^[1] Multiple factors can affect the livelihoods of individuals or the capacities of institutions. For policy makers as well as for stakeholders it is important to know what the added value is of the policy intervention apart from these other factors. The attribution problem is often referred to as the central problem in impact evaluation. The central question is to what extent can changes in outcomes of interest be attributed to a particular intervention?^[2] Attribution refers both to isolating and estimating accurately the particular contribution of an intervention and ensuring that causality runs from the intervention to the outcome. In most contexts, adequate empirical knowledge about the effects produced by an intervention requires at least an accurate estimate of what would have occurred in the absence of the intervention (the counterfactual) and a comparison with what has occurred with the intervention implemented.

Box 1 – The attribution problem

Analyzing attribution requires comparing the situation ‘with’ an intervention to what would have happened in the absence of an intervention, the ‘without’ situation (the counterfactual). Such comparison of the situation ‘with and without’ the intervention is challenging since it is not possible to observe how the situation would have been without the intervention, and has to be constructed by the evaluator. The counterfactual is illustrated in the Figure below. The value of a target variable (point a) after an intervention should not be regarded as the intervention’s impact, nor is it simply the difference between the before and after situation (a-b, measured on the vertical axis; the dotted arrow). The net impact (at a given point in time) is the difference between the target variable’s value after the intervention (a) and the value the variable would have had in case the intervention would not have taken place (c).

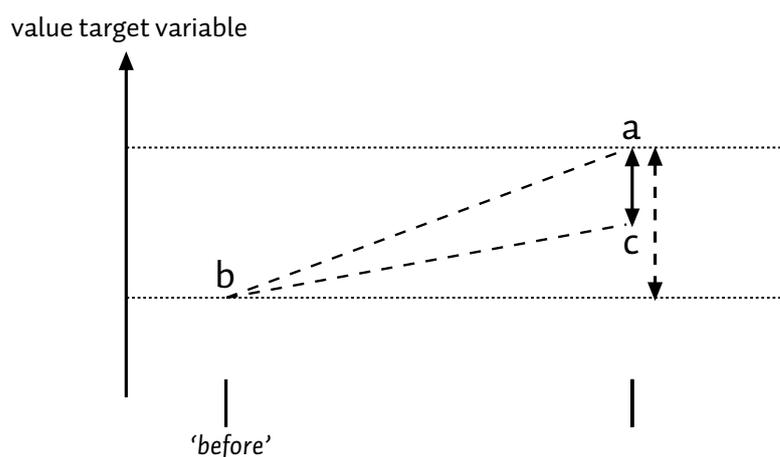


Figure 2 - Graphical display of the net impact of an intervention

[1] Goal achievement can be assessed without the need for attribution analysis as no differentiation is made between whether changes are due to the intervention or other factors.

[2] Attribution is about causal relationships between intervention and effects. How and to what extent the attribution challenge has been addressed in an impact evaluation determines the internal validity of findings.

Usually interventions target particular groups, and in addition self selection effects may occur as more motivated or socially well-positioned individuals or groups gain access to a particular program. Consequently, one cannot simply compare the situation of participants over time with the population at large. Estimates would be distorted due to this selection bias problem. The safest way to avoid selection effects is a randomized selection of intervention group and control before the experiment starts. When the experimental group and the control group are selected randomly from the same eligible population, in sufficiently large samples both groups will have similar average characteristics (except that one group has been subjected to the intervention and the other not). Consequently, in a well-designed and correctly implemented RE a simple comparison of average outcomes in the two groups can adequately resolve the attribution problem and yield accurate estimates of the net effect of the intervention on a variable of interest: by design, the only difference between the two groups was the intervention.

This powerful feature of REs explains the increasing popularity of this methodology. With a long tradition in medicine and public health and a much younger tradition in policy fields such as education and crime and justice (Leeuw, 2009), REs are now also increasingly applied in the context of (social) development interventions. Randomization is not always feasible, but a wide variety of quasi-experimental designs are available to ensure a high internal validity of findings. Basically, designs differ in terms of the technique used for creating comparable groups (e.g. regression discontinuity, propensity score matching, pipeline approaches) as well as in terms of the structure of periodic measurement within participant and control groups (e.g. simple ex ante – ex post participant group design, interrupted time series design; see for example Campbell, 1969; Cook and Campbell, 1979; Shadish et al., 2002, Morgan and Winship, 2007; for development interventions see for example Bamberger et al., 2006).^[1]

4.2. Internal versus external validity

REs are typically equipped to address the question of what works within the particular confines of the experiment; they are strong on internal validity. However, policymakers are often interested in other questions (Heckman, 1992; Heckman et al., 1997; Ravallion, 2009b). Does this intervention also work in other regions or contexts? What happens when we go to scale with a particular intervention? What are the determinants of effectiveness? Another typical question that might not be easily answered with REs or quasi-experiments, is whether and how people are differently affected by an intervention.^[2] This question can be answered with additional quantitative data analysis, if (large) data sets are available which allow for extensive modeling of confounders and interaction effects. Alternatively (or in addition), many qualitative methods such as case study methods can help evaluators to study in detail how interventions work differently in different situations.

[1] Most quasi-experimental techniques are useful when selection characteristics are known and can be measured. Even in the case of unobservable characteristics which might differ between groups and affect intervention effects, this may not be a problem. If these characteristics are time invariant, in principle they can be controlled for by double differencing or multiple data points in time.

[2] This is an issue that is closely related to the idea of external validity. If one knows how an intervention affects groups of people in different ways, then one can more easily generalize findings to other similar settings.

Without further information results of a RE cannot be generalized beyond the experimental setting, as important confounders that are controlled for in the experiment are not revealed by the experiment itself. Moreover, “[t]he people who are normally attracted to a program, taking account of the expected benefits and costs to them personally, may differ systematically from the random sample of people who were included in the trial” (Ravallion, 2008: 17). Critics of the ‘alleged superiority’ of REs argue that internal validity is not typically the question that policy makers are interested in and argue for more attention to external validity of findings, an aspect on which REs enjoy no comparative advantage (e.g. Rodrik, 2008; see Shadish et al. (2002) for a discussion on experiments and external validity). They argue that in order to be able to generalize findings from a RE to other settings, one needs to know how an intervention works, what are the determinants of the processes of change (possibly induced by an intervention), how an intervention might affect people in particular circumstances in different ways and what the time path and nature of the changes might be. In order to answer these types of questions (and others, see Ravallion, 2009b) one needs an informative explanatory theory (e.g. based on research within the social sciences) and other (additional) methods of data collection and analysis (e.g. Van der Knaap et al., 2008; for the case of development interventions see Deaton, 2009).

‘Randomistas’ have asserted that external validity of findings can be enhanced by doing a series of experiments in different contextual settings (Banerjee and Duflo, 2008). Yet as Ravallion argues, “the feasibility of doing a sufficient number of trials – sufficient to span the relevant domain of variation found in reality for a given program, as well as across the range of policy options – is far from clear. The scale of the randomized trials needed to test even one large national program could well be prohibitive” (Ravallion 2008: 19). Moreover, as Rodrik argues, “[f]ew randomized evaluations – if any – offer a structural model that describes how the proposed policy will work, if it does, and under what circumstances it will not, if it doesn’t. Absent a full theory that is being put to a test, it is somewhat arbitrary to determine under what different conditions the experiment ought to be repeated.” (Rodrik, 2008: 21-22; italics added). A similar point is made by Deaton who asserts that “repeated successful replications of a ‘what works’ experiment is both unlikely and unlikely to be persuasive. Learning about theory, or mechanisms, requires that the investigation be targeted towards that theory, towards why something works, not whether it works” (Deaton, 2009: 31).

A special type of generalizability concerns the inference from small-scale pilots to up-scaled programs with broad coverage. In development contexts REs are often implemented on small scale pilots. While the outcomes of REs are considered as useful inputs in the decision to go to scale, REs alone are insufficient to support such a decision.^[1] In case of scaling-up, conditions at both the institutional level (of implementing agencies) as well as the beneficiary level invariably change, making the new intervention altogether different from the small-scale pilot it was derived from (Deaton, 2009). At the institutional level for example, the organizational set-up might change completely, with new people with divergent capacities and interests managing and implementing the new intervention, or corrupt public officials suddenly being drawn to a scaled-up intervention which has appeared on their radar. In addition, new contextual conditions and characteristics of target groups may confound intervention effects, this new heterogeneity not being covered by the original small-scale experiment (Ravallion, 2009a). On the other

[1] In fact, there are good examples of REs convincing policymakers to scale up and replicate interventions. For example, the spread of conditional cash transfer programs over Latin America is in part fuelled by the evidence produced by REs.

hand, it can be argued that a RE which has covered a sizeable sample can be quite informative on the likelihood that the same policy instrument will produce similar results when scaled up within the same population the RE sample was taken from.^[1]

4.3. Interventions as theories

Most critics of REs subscribe to the point of view that there is nothing that precludes REs from being more theory-based or -driven. Of course, in such cases it would be the combination of theory (or theories) and RE that would increase the external validity (and also internal validity) potential of REs rather than the RE format alone. Recently, more examples of theory-based REs can be found in the literature (see for example Banerjee (2005) or Banerjee and Duflo (2008) for illustrations).^[2] Moreover, as argued by Cook (2006) REs are never completely theory-empty.^[3] Among other things, they require substantive theory as guidance for selecting or constructing suitable measures of effects.

While REs in principle are not or need not be theory-empty, the abovementioned critique of Deaton and others on REs remains valid. REs are geared towards the question of whether an intervention works and do not shed much light on how and why interventions work. In order to look at the latter question, evaluation researchers need to look into the black box between output, outcome and impact indicators and reconstruct the underlying causality (Pawson and Tilley, 1997). Two types of theory are of importance here. A reconstructed intervention theory should adequately represent the main assumptions of decision makers, target groups and other stakeholders about causal pathways running from intervention to effects (see for example, Chen, 1990, Rogers et al., 2000; Leeuw, 2003). The intervention theory can be further enriched or tested by taking into account existing explanatory theories (from the social and behavioral sciences) on intervention contexts, processes of change and potential effects.

The idea that existing explanatory theories can improve the quality of evaluations is not a new one and has been discussed quite extensively in the literature (e.g. Riggin, 1990; Lipsey, 1993; Donaldson and Lipsey, 2006). For example, Riggin (1990) discusses how Etzioni's theory of compliance can improve the quality of an evaluation of an employment assistance program both at the stage of conceptualization as well as at the interpretation stage of an evaluation. A first important role for 'theory' lies in the conceptualization of an evaluation question. Substantive theories help to point the evaluator toward the relevant constructs and relationships between these constructs in order to make useful abstractions of the reality of a policy intervention, its intended effects and the wider context in which it is embedded, aspects which subsequently can be tested through empirical research. Substantive theory from past (evaluation) research to some extent can help to anticipate these effects, which subsequently can be taken into account by means of additional data collection. A second important role for theory

[1] Thereby assuming that first a random sample was taken from the population, with subsequent randomized allocation of the intervention to treatment and control groups within the sample.

[2] One example is Duflo and Hanna (2008) on using an experiment to develop a model of teacher behavior.

[3] Theory-empty refers here to a situation in which the causal relationships between intervention outputs, outcomes and impact are not made explicit in a theoretical framework. Such a framework may be based on reconstructions of stakeholders' assumptions and/or existing research relevant to the causal pathways in question. For example, an analysis in which a claim to a relationship or even a causal relationship between two variables is based on statistical association only (whether in an experimental setting or not), can be called theory-empty (for a discussion see for example Coleman, 1986).

lies in reinforcing the causal analysis, the analysis of how and to what extent changes in target variables can be attributed to a policy intervention. Relevant substantive theories can shed light on the nature of the causal relationships between interventions and (un)intended processes of change and help to rule out rival explanations for changes in target variables.^[1] Finally, there is an important role for theory in the interpretation of evaluation findings. Theory can provide a useful framework for helping us to understand why certain changes have come about or provide insights into the relevant (contextual) variables which are likely to influence patterns of results across settings.

4.4. The law of comparative advantages: theory-based and multi-method evaluation

My account so far has gradually led us the realization that REs are potentially strong on internal validity yet miss out on providing (strong) evidence on other aspects such as for example unanticipated and long-term effects and the external and construct validity of findings. This brings us to the important realization that impact evaluations by default cannot and should not exclusively rely on one methodological design only.^[2] The remainder of this section will add some more illustrative power to this argument.

As discussed elsewhere (Leeuw and Vaessen, 2009), the intervention theory constitutes the guiding framework of impact evaluations. Typically, multiple causal assumptions emerge that require further testing in order to be able to claim whether the intervention has induced certain changes, and in what circumstances. However, not all causal assumptions can be tested with the same methodological design or specific method. For example, consider the effects of subsidies for sustainable land use on biodiversity. One can envisage a useful RE which tests the effects of subsidies on the land use behavior of farmers. The subsequent causal step from land use behavior to biodiversity cannot be tested so easily by means of a RE. One of the main reasons is that biodiversity depends on plot-specific variables (e.g. the type of vegetation on a certain plot of land) as well as on determinants at higher levels such as the connectivity between systems of land use, the proximity of certain biospheres, the geographical location with respect to migration routes of birds and other species, and so on. Moreover, as argued above, there are other challenges such as the non-linearity, uncertainty and sustainability of changes.

Ideally, the intervention theory should provide guidance on the choice of causal assumptions to be analyzed (see Weiss (2000) for a discussion on this) and correspondingly, different designs and methods can be used to assess specific assumptions. In addition, other complementary perspectives on the use of multiple methods can be discerned in the literature (for a general discussion see for example Tashakkori and Teddlie, 2003).

[1] An example of such an approach is Scriven's (2008) General Elimination Methodology. See also Pawson's (2006) discussion on adjudication between rival theories. In an evaluation of the impact of social funds, Carvalho and White (2004) reconstruct a theory and an 'anti-theory', which are both put to the test empirically, in order to arrive at a better understanding of the nature of change processes.

[2] "[I]t is trivial to argue about whether evidence-based research should be multi-method or not. Even causal research is deepened by learning about non-causal issues, such as what the substantive theory behind a program design is, who gets to participate in it or not, how well the program is implemented, who gets greater exposure, and what the program costs. So nearly all causal studies require multiple methods that complement each other. Multi-method, complementary research is desirable even when a causal claim is centrally at issue" (Cook, 2006: 1).

Let us briefly illustrate three ‘logics’ for multi-method approaches in impact evaluation. The first starts out from Campbell’s framework of different types of validity. As suggested earlier in this paper, specific methods have comparative advantages in ensuring a high degree of internal/external/construct validity. Consider a similar example as introduced above, i.e. an intervention that provides monetary incentives and training to farmers in order to promote land use changes leading to improved livelihoods conditions as well as other effects. Simplified, a comprehensive methodological design could be the following:

E.g. a randomized experiment with the use of survey data on participant and control groups could be used to assess the effectiveness of different incentives on land use change and/or socio-economic effects of these changes (potentially strengthens internal validity of findings);

E.g. further survey data (multivariate) analysis and case studies could tell us how incentives have different effects on particular types of farm households (potentially strengthens internal validity and increases external validity of findings);

E.g. targeted syntheses of existing research, semi-structured interviews and focus group conversations could tell us more about the nature of effects in terms of production, consumption, poverty, environment etc. (potentially enhances construct and internal validity of findings).

A second framework of multi-method evaluation has been labeled the ‘shoestring’ approach (see Bamberger et al., 2004). It refers to a number of scenarios of multi-method approaches which are adapted to particular conditions of budget, time and data constraints. These scenarios all rely on an intervention theory model as a basis for different methodological strategies to simplify evaluation design (e.g. in relation to text book REs), reduce costs of data collection and analysis, and integrate qualitative and quantitative methods.^[1]

A third illustration of a multi-method perspective is Woolcock’s account of different options to gain a better understanding of the nature of causal pathways. “There are at least three entry points, each of increasing degrees of sophistication. The first is simply raw experience: seasoned project managers should have a good sense of how long and in what form the impacts associated with a particular project in a particular context should take to materialise. [...] Astute intuition and seasoned field experience combined with solid theory should provide a second avenue: the very essence of a good theory should be that it provides a sense and a justification of the conditions under which, and mechanisms by which, certain project outcomes should be expected. [...] Thirdly, the regular collection of empirical evidence can itself be a basis for determining the shape of the project’s impact trajectory, and is ultimately (for researchers at least) the most defensible basis on which to do so” (Woolcock, 2009: 7-8).

[1] See also Bamberger et al. (2009) for further discussion on mixed methods in the context of impact evaluation.

5. CHALLENGE 3: IMPACT EVALUATION IN PRACTICE

5.1. Threats to attribution analysis in experimental settings

The main threats to the internal validity of findings from REs (and certain quasi-experiments) are well-known and widely discussed (Campbell and Stanley, 1963; Campbell, 1969; Cook and Campbell, 1979; Shadish et al., 2002):

Selection bias: refers to the problem of under- or overestimating intervention effects due to uncontrolled differences between different (treatment) groups that would lead to differences in effect variables if none of the groups would have received benefits.

Contagion (or treatment diffusion): refers to the problem of groups that are not supposed to be exposed to (or receiving) certain benefits are in fact benefiting from an intervention in one or more ways: by directly receiving the benefits from the intervention, by indirectly receiving benefits through other participants, or by receiving similar benefits from other organizations.

Aging/Maturation: An effect that arises when participants grow older, wiser, more tired, more self-confident due to factors other than the intervention.

History: The effect is caused by some event other than the intervention occurring during the same time period as the intervention.

Testing: The pre-test measurement causes a change in the post-test measure.

Instrumentation: The effect is caused by a change in the method of measuring the outcome.

Regression to the Mean: Where an intervention is implemented on units with unusually high scores (e.g. unusually high student performance scores), natural fluctuation will cause a decrease in these scores on the post-test which may be mistakenly interpreted as an effect of the intervention.

Attrition: Changes in effect measurements due to drop-outs.

Causal Order: It is unclear whether the intervention preceded the outcome.

Behavioral responses: Several unintended behavioral responses not caused by the intervention or 'normal' conditions might inhibit the reliability of comparisons between groups and hence the ability to attribute changes to the intervention. An example is expected behavior or compliance behavior: beneficiaries' behavior is not caused by intervention outputs (e.g. a subsidy) but due to reasons of compliance with a (formal/informal) contract between beneficiary and implementing organization, due to the (longstanding) relationship with a particular organization (delivering intervention outputs), or due to certain expectations about future benefits from the organization (not necessarily the intervention).

While each of these issues might be potential problems that render claims on attribution less valid, the fact that they have been systematically identified and addressed in the literature (see for example Cook and Campbell, 1979) enhances the scientific rigor of experimental and quasi-experimental designs.

A key underlying determinant of the internal validity of findings from REs is the extent to which those managing the experiment are capable and willing to safeguard the design from the threats to validity described above. This is certainly the case for selection bias and treatment diffusion issues; well-designed REs may be safeguarded from these problems. More complicated is the potential threat from unintended behavioral responses among participant or control groups. As argued by several authors (e.g. Deaton, 2009; Cohen and Easterly, 2009)

randomization can lead to hard feelings between the different groups involved in the experiment. “[S]ubjects may fail to accept assignment, so that people who are assigned to the experimental group may refuse, and controls may find a way of getting the treatment, and either may drop out of the experiment altogether. The classical remedy of double blinding, so that neither the subject nor the experimenter know which subject is in which group, is rarely feasible in social experiments” (Deaton, 2009: 36). In some cases researchers cannot (and indeed should not) withhold the information from stakeholders that they are part of an experiment. Banerjee and Duflo (2008) provide several examples of mechanisms (e.g. lotteries) which facilitate the collaboration of local populations in an experiment. Yet, even in situations where institutions and target groups agree to randomized allocation of benefits, it is well-known in the literature (e.g. Campbell, 1969; Shadish et al., 2002) that experimentation may lead to a range of unintended behavioral responses within the participant and treatment groups which can affect the validity of findings resulting from an experiment. An example of unintended behavioral responses comes from the famous PROGRESA study (see for example Skoufias and McClafferty, 2001). “One issue with the explicit acknowledgement of randomization as a fair way to allocate the program is that implementers will find that the easiest way to present it to the community is to say that an expansion of the program is planned for the control areas in the future (especially when it is indeed the case, as in phased-in designs). This may cause problems if the anticipation of treatment leads individuals to change their behavior. This criticism was made in the case of the PROGRESA programs, where control villages knew that they were going to eventually be covered by the program” (Banerjee and Duflo, 2008: 22).

5.2. Other design and implementation challenges

A key issue regarding the success of REs in practice revolves around the ethics and feasibility of randomization in practice and the corresponding reactions by stakeholders. Experiments can cause resentment as people do not understand or support the differences in benefits allocated to different groups. Random allocation in many cases is also unacceptable to policy makers (e.g. Bamberger and White, 2007; Ravallion, 2008). Interventions are often intended to be targeted to specific groups and outreach is a direct concern to policy makers. Randomization would limit outreach and at the same time is often considered as unethical in view of the pressing needs of target populations.^[1]

Banerjee and Duflo (2008: 22) argue that collaboration with institutions in developing countries is becoming less of an issue in developing countries. “Randomization that takes place at the level of location can piggy-back on the expansion of the organization’s involvement in these areas limited by budget and administrative capacity, which is precisely why they agree to randomize. Limited government budgets and diverse actions by many small NGOs mean that villages or schools in most developing countries are used to the fact that some areas get some programs and others do not and when a NGO only serve some villages, they see it as a part of the organization’s overall strategy. When the control areas [are] given the explanation that the program had only enough budget for a certain number of schools, they typically agree that a lottery was a fair way to allocate it---they are often used to such arbitrariness and so randomization appears transparent and legitimate”.

Another way to enhance the goodwill among implementing agencies is to forge a

[1] This often provides a compelling argument for using quasi-experimental designs. For example, regression discontinuity is very useful when targeting is based on clear and easy to measure criteria of selection.

longstanding relationship between the latter and RE researchers in which a series of experiments will constitute the basis of a cumulative process of learning. It still remains to be seen however what the costs and benefits of such a relationship would be in divergent institutional contexts, especially in view of the previously discussed issues. Institutional willingness to undertake a RE is not a black and white issue, as interventions typically comprise multiple institutional partners and multiple layers of management, from headquarters to field level. The idea of institutional willingness is also closely related to institutional capacities and incentives. REs are intrinsically linked to intervention implementation processes and the question to what extent well-trained researchers are de facto present and able to ensure the quality control of experimental conditions, is an empirical one that merits further inquiry. There is a marked difference between an experienced research team undertaking REs (as is the case for most of the published work on REs in development) and the idea of mainstreaming REs in the design of selected projects of donor and developing country agencies' portfolios. In the latter case invariably not all the tasks pertaining to the design and implementation of a RE are managed by experienced and motivated researchers.

Apart from institutional capacities, willingness to collaborate and ethics, other challenges remain. A first obvious condition for REs is the active involvement of researchers or evaluators in the intervention design and implementation phase. This involvement is essential for baseline data collection as well as quality control of randomization. In practice however, many impact evaluations are commissioned after an intervention has been implemented and baseline data continue to be a problem (Bamberger and White, 2007). Although preferably double difference (participant-control group comparisons over time) designs should be used, it is more usual that impact assessments are based on less rigorous – and reliable – designs, where baseline data are reconstructed or collected later during the implementation phase, or baseline data are collected only for the treatment group, or there are no baseline data for neither treatment nor control group (for options on how to address these constraints see Bamberger et al., 2004; Bamberger et al. 2006; or White and Bamberger, 2007).

A second aspect is the costs of doing REs. Early experiences of REs (and quasi-experimental studies) in development by the World Bank turned out to be rather costly (OED, 2005), but these studies were often quite ambitious in scope. More recent experiences seem to suggest that REs do not necessarily have to be more expensive than other non-experimental observational studies that are based on original fieldwork with multiple data points in time. However, given the narrow focus of REs, when large programs with multiple intervention activities need to be evaluated, REs need to compete with less expensive non-experimental methodologies which can cover a broader scope of activities with less budget. Proponents of REs will need to justify if the potential gains in terms of the high internal validity of findings delivered by an RE will be worthwhile the investment given the loss in scope. Alternatively, adversaries of experimentation or others that choose scope over depth will have to justify that alternative methodological designs adequately address crucial issues such as attribution.

A third issue concerns the quality of the data. Bamberger and White (2007) argue that problems of data quality, although relevant for any type of methodological design, might be particularly problematic in case of REs as they rely on a limited number of indicators. Banerjee and Duflo (2008) contest this idea. In their view, if data is being collected especially for

the purposes of a RE, then given the limited number of observations that are usually necessary for reliable estimates, researchers are able to dedicate special attention to data collection and measurement issues. Ensuring high data quality is particularly challenging in rural contexts in developing countries.^[1] Surveys continue to be the main instruments generating impact evaluation data. Potential measurement errors due to problems of recall, sensitivity of certain topics, intercultural communication, translation errors are just a few of the problems that affect data quality (see for example De Leeuw et al., 2008; Mikkelsen, 2005; Bamberger et al., 2004; Bamberger, 2000). Moreover, surveys may not be appropriate for addressing sensitive topics (see for example Bamberger et al., 2009) such as for domestic violence, household income or local norms and in these cases are more likely to generate unreliable data. Unfortunately, data quality is not as sexy a topic as methodological design, especially if the latter is the acclaimed basis for delivering 'rigorous scientific evidence'.

[1] For a rather critical perspective on data quality in (rural) developing country contexts, see Gill (1993).

6. SOME LESSONS FOR RANDOMIZED EXPERIMENTS AND IMPACT EVALUATION IN GENERAL FROM THE PERSPECTIVE OF A ‘NON-RANDOMISTA’

The controversy surrounding the promotion and application of REs in development has led to a sense of polarization in the development policy and evaluation community. As some proponents claim epistemological supremacy of REs (with respect to attribution) the counter reaction among others has been rejection. Needless to say, such extreme positions are counter-productive to reaching a goal that is commonly endorsed: to learn more about what works and why in development. Polarization leads to ‘argument mining’, with proponents bringing up (arguably valid) arguments in defense of REs while adversaries pick their favorite (and arguably valid) arguments against REs. Clearly, this is not the way forward on the path to knowledge growth. If one explores the growing body of evidence in development generated through REs, one cannot deny the positive (though divergent) direct and indirect benefits to knowledge generation about what works. At the same time, as acknowledged by most scholars in the literature, the applicability of REs is limited to certain contexts. Moreover, as illustrated previously, the range of potential challenges (including different questions) to be addressed in an empirical exercise of assessing what works and why in development intervention is too broad to be adequately addressed by REs only. By presenting a diverse array of methodological and conceptual perspectives on impact evaluation this paper has illustrated potential benefits but also limitations as well as complementary perspectives to REs. The implicit lesson is twofold. First of all, the question ‘to randomize or not to randomize’ is overrated in the current debate. Other challenges demand the attention of policymakers and evaluation researchers alike. Second, ‘do not throw out the baby with the bath water’. There is a risk that the current popularity of REs in certain policy circles might lead to a backlash. Too high expectations of REs may quicken its demise.

An important barrier to RE application is that policy-driven impact evaluations, i.e. impact evaluations commissioned by funding or implementing agencies, often favor scope over depth. With a limited budget, evaluators are forced to develop plausible statements on impact over a range of intervention activities in divergent contexts. In such cases, a tension between accountability (implying a coverage of most or all of the funded activities) and learning (giving more attention to one particular intervention or type of intervention) may exist.^[1] Even if the applicability range of REs broadens in the future due to new experiences (e.g. experiments at the institutional level), aspects such as implementation challenges (e.g. the ethics, logistics of doing REs but also the willingness among policy makers, implementing agencies and other stakeholders to transform intervention formats into randomized experiments), budget allocation priorities (e.g. scope versus depth) and other limitations will inevitably restrict RE-type evaluations to being a minority practice in the impact evaluation business (see below).

The realization that there are inherent limitations in applicability of REs should not be an argument against the further promotion of REs. Nor should the range of threats to validity that can affect the analytical strength of an RE constitute an argument against REs.^[2] I

[1] See CGD (2006) for a broader discussion on the incentive problems that explain the limited investments of donors in REs.

[2] In fact, the systematic identification and discussion in the literature of threats to the (internal) validity of REs strengthens the scientific basis of the methodology. As such, it should not be considered as an argument against REs. However, caution is in order when talking about the possibility of mainstreaming REs in the intervention cycle of an agency. I remember a recent debate in a multilateral donor organization where there was talk of mainstreaming REs in the design of projects. It is very likely that such experiments will not benefit from the same level of quality control

briefly discuss a few points of interest here. First, experimentation should not be perceived as an anomaly in development intervention. The policy field of education in the Netherlands is in this aspect not much different from the policy field of agricultural technologies in Latin America. From the former perspective, Borghans (2009) argues that teachers continuously experiment with teaching methods, whereas the target group, students, are continuously receiving different ‘treatments’, for example simply by going to different schools. Similarly, in Latin America (or elsewhere), farmers and development agencies alike continuously experiment with new techniques and new intervention activities to achieve particular desired objectives.^[1] If one considers a specific sample of farmers or agencies, at any given point in time, one can identify a range of similar yet different practices aimed at boosting productivity per unit of land given particular resource constraints. Observation over time in combination with a systematic variation in respectively the application or promotion of certain techniques, are common behavior among farmers, cooperatives, NGOs or state agencies. A RE is a more systematic approach to experimentation than what most stakeholders are used to. As argued by Borghans (2009) from the perspective of education in the Netherlands, if we take this experimenting attitude as a given, then one would expect it to become more acceptable and desirable for decision makers and target groups to organize and participate in more systematic experiments, such as REs. In the context of development interventions Banerjee and Duflo (2008) talk about developing long-term working relationships between researchers and development agencies. In such conditions it is possible to show that a RE is much more efficient in proving whether something works than much of the haphazard experimentation that goes on in the daily practice of target groups and implementing agencies. And it is in such cases that the comparative advantage of a RE can truly blossom: it magnifies heterogeneity in ‘treatment’ by introducing a clear comparative perspective between those with and without an intervention and it reduces bias in estimation of effects through the principle of randomization. These are two powerful features that help us to identify efficiently what works for a particular group in a particular context.

Now, in order to go from a specific setting to a more generalizable conclusion about effectiveness we need theory. To illustrate this, consider the adoption of certain organic farming practices. We know from research that knowledge and labor substitute for capital. In other words, less inputs are bought on the market in exchange for an increased input of labor and knowledge. Peasant economics (e.g. Ellis, 1988) and the diffusion of innovation literature (e.g. Rogers, 2003) teach us that the opportunity costs of labor (next to a range of other variables, and contingent upon among other things the type of crop) is an important explanatory variable of smallholder adoption behavior. With this information in mind we can test whether this theory holds for a specific farmer population or region. For example, we may set up a clustered RE with different samples in several regions which mainly differ in the opportunity costs of labor yet are similar in other potential explanatory variables (as derived from theory). While this example may not be water tight, it can prove to be very informative on the effectiveness of certain policy instruments (e.g. subsidies, training) in promoting the adoption of organic practices and at the same time test whether the opportunity costs of labor is a decisive factor in adoption behavior. Of course, the stratification may be a little bit more complex as more explanatory variables (derived from theory) may determine the setup of the series of REs. The core idea is that in this way REs may more effectively contribute to existing theories on the determinants of adoption

present in most (research-driven) RE studies which feature so prominently in the literature.

[1] However, smallholder farmers are often risk averse and tend to experiment on a small scale. Once a particular new practice has demonstrated its pay-off, experimentation may lead to broader adoption.

processes. In general, it shows that theory may be rigorously tested by means of REs and that theory itself can be a guideline for determining how to set up a series of REs (see also Cohen and Easterly, 2009, for a discussion on the theory-testing potential of REs). Limitations to external validity remain, especially with respect to scaling-up effects. Nevertheless, a theory-driven RE is potentially stronger on external validity than a ‘theory-empty’ RE. As commented by Deaton (2009) and also Banerjee and Duflo (2008) the number of theory-informed and theory-testing REs is on the increase. This is important as REs without a basis in theory are prone to be weaker not only in terms of external validity, but also the internal validity and construct validity of findings.

The biggest gains from this growing attention for rigorous impact evaluation, this ‘new push for objective evidence’ as one may call it, are not to be found in the growing body of REs but rather in its spin-offs. Whereas development evaluations used to be largely on process issues or output delivery, the paradigm shift in development policy towards more attention for results has strengthened the belief across the developing world that interventions should not be based on hunches, intuitions or ideologies but on evidence on whether an intervention is likely to make a difference in terms of the desired objectives. A randomized experiment in a way is one of the elegant flagships of this new evolution. As a methodological design it has drawn a lot of attention yet it is unlikely to win the war on its own. Several other trends and opportunities can be noted which in part have been strengthened by the ‘randomista’ movement. These (partial) ‘spin-offs’^[1] point at other challenges in impact evaluation. First of all, a RE is just one of the designs based on explicit counterfactual analysis. The number of quasi-experimental studies in development has increased sharply over the last ten years or so. For example, where randomization is not possible or appropriate, regression discontinuity analysis may be used instead, as it relies on a different principle for the definition of groups.^[2] In addition, as data gathering efforts have increased under the influence of results-based thinking and new M&E systems, opportunities for ex post statistical matching or multivariate analyses with statistical controls have also markedly increased.

This paper has focused on design and implementation issues in impact evaluation, scarcely touching upon the large and expanding body of statistical impact evaluations. Statistical techniques are often used within the framework of (quasi-)experimental designs but more often than not in non-experimental settings. Comparative advantages of (non-experimental) statistical impact evaluations are among other things their broad coverage in terms of the number of individuals, households, districts, and regions encompassed by the data sets. The increased availability of data and the growing capacities (in terms of expertise and technological support) to process and analyze these data have enhanced the opportunities for impact evaluation at aggregate (e.g. regional) levels. The issue of the comparative advantage of REs vis-à-vis non-experimental statistical analyses has been briefly raised in this paper. One particular argument all too often is ignored. As a methodological design requiring active manipulation of an intervention, an RE relies on original data, collected specifically for the purposes of the study. In most cases, the evaluation researchers analyzing the data are also involved in the data collection. In qualitative research one also commonly finds a strong link between data collection and analy-

[1] In fact, it is better to think in terms of an association between growth in REs and other quantitative methodological designs.

[2] Comparing over time a participant group beneath a certain threshold value of a particular targeting variable (e.g. distance to road) with a control group just above the threshold value.

sis. Not necessarily so in statistical impact evaluation exercises. All too often data analyses are based on data sets constructed by others for other purposes. Econometricians and economists involved in statistical impact evaluations using (mostly) non-experimental data tend to overemphasize ‘threats to validity’ in the data analysis phase (e.g. specifying the right selection model) and often blatantly ignore (or are ignorant of) any problems or biases that may have arisen in the data collection phase. The severed link between data collection and analysis in many impact evaluation exercises is distressing and merits a higher profile in methodological debates.

Data quality is also a concern in the literature on mixed methods in impact evaluation.^[1] Previously, I already underlined the importance of mixed methods from the perspective of comparative advantages of methods. Two other reasons make this area of research particularly relevant to policymakers and evaluation researchers. First of all, practically all evaluation work is multi-method in nature. This is most clearly visible in large program or portfolio evaluations where approach papers and evaluation matrices specify the range of data collection and analysis tools used to analyze specific questions. A second reason is that the majority of impact evaluations take place under less than ideal circumstances (see Bamberger et al., 2009). Often evaluation researchers are not present in the design phase of interventions; they are called in when an intervention is already in the implementation phase or after completion. Consequently, evaluators often have to resort to baseline reconstruction, secondary data or ex post data only. In addition, budgets often do not permit large sample sizes or elaborate designs with multiple group comparisons. Pressures on scope further limit the chances of evaluators to set up for example a solid quasi-experiment. Time pressures may force evaluators to do the work in ‘quicker and dirtier’ ways than what is needed to appropriately address the attribution issue. Methodological options for mixed method evaluations under less than ideal circumstances are existent yet, as argued by one of the principal authors on this subject, Michael Bamberger, much remains to be done in terms of developing new methodologies and standards for mixed method evaluations (see for example Bamberger et al., 2004, 2009).

If decision makers want answers to their questions on what works across interventions and across contexts, they may need to change part of their focus. Instead of seeing administrative categories such as projects or country portfolios as the only principal units of analysis (often mainly for accountability purposes) they need to look more at particular policy instruments or intervention types which recur throughout projects.^[2] Consequently, decision makers may learn to appreciate and invest more in rigorous evaluations of particular intervention types (e.g. using REs) and see how these pieces of evidence connect to the macro picture, i.e. looking at the importance of a particular intervention type and the divergent contexts in which it is implemented. Such a focus on particular intervention types and policy instruments can perfectly coexist with more comprehensive approaches to impact evaluation which start out from programs and portfolios, including a variety of strategies and interventions.

[1] For examples see Leeuw and Vaessen (2009) or Bamberger et al. (2009). Karlan (2009) argues the case for more mixed method research in an RE context.

[2] For example, even though microcredit may only be a small component of a particular project or country portfolio of a donor agency, throughout its entire portfolio a sizeable portion of the budget may be allocated to supporting microcredit activities (at different levels).

We know there are no laws in social sciences (Elster, 2007), yet we also know there patterns of regularity, or demi-regularities, in individual, social and institutional behavior (Pawson, 2009). Identifying and refining such patterns of regularity require a particular way of theorizing about interventions, deconstructing or unpacking interventions into their active ingredients: policy instruments linked to contexts linked to behavioral mechanisms (see Pawson, 2006, 2009). It is our mission as development researchers and evaluators to uncover these patterns of (demi-)regularities as they are the building stones of knowledge about what works and why across interventions. Intervention theories with a certain degree of external validity are becoming more and more important in the context of review and synthesis work as well. 3IE is one of the organizations which is currently investing in this type of work (for an example on microcredit see Vaessen et al., 2009).^[1] We still have a long way to go. One thing is for sure, we need good empirical impact evaluations – which rely on intervention theory, explanatory theory and multiple methods tailored to a specific context – in order to succeed, eventually...

[1] Comparable to the work by institutions such as the Campbell and the Cochrane Collaboration on health, education, crime and justice and social work, based on empirical work from (mostly) OECD countries.

REFERENCES

- Baker, J.L. (2000) *Evaluating the impact of development projects on poverty*, World Bank, Washington D.C.
- Bamberger, M. (2000) "Opportunities and challenges for integrating quantitative and qualitative research", in: M. Bamberger (ed.) *Integrating quantitative and qualitative research in development projects*, World Bank, Washington D.C.
- Bamberger, M. and H. White (2007) "Using strong evaluation designs in developing countries: Experience and challenges", *Journal of Multidisciplinary Evaluation* 4(8), 58-73.
- Bamberger, M., J. Rugh, M. Church and L. Fort (2004) "Shoestring Evaluation: Designing impact evaluations under budget, time and data constraints", *American Journal of Evaluation* 25(1), 5-37.
- Bamberger, M. J. Rugh and L. Mabry (2006) *Realworld evaluation: Working under budget, time, data, and political constraints*, Sage Publications, Thousand Oaks.
- Bamberger, M., V. Rao and M. Woolcock (2009) "Using mixed methods in monitoring and evaluation: Experiences from international development", *MIMEO*, World Bank, Washington D.C.
- Banerjee, A. (2005) "New development economics and the challenge to theory", *Economic and Political Weekly* 40, October 1-7, 4340-4344.
- Banerjee, A.V. and E. Duflo (2008) "The experimental approach to development economics", *NBER Working Paper* 14467, Cambridge.
- Banerjee, A., S. Cole, E. Duflo and L. Linden (2007) "Remedying education: Evidence from two randomized experiments in India", *Quarterly Journal of Economics* 122(3), 1235-1264.
- Blattman, C. (2008) "Impact Evaluation 2.0", *Presentation to the Department of International Development*, February 14, 2008, London.
- Borghans, L. (2009) "Leren over leren", in: R. Rouw, D. Satijn and T. Schokker (eds.) *Bewezen beleid in het onderwijs*, Ministerie van Onderwijs, Cultuur en Wetenschap, The Netherlands.
- Campbell, D.T. (1969) "Reforms as experiments", *American Psychologist* 24, 409-429.
- Campbell, D.T. and J.C. Stanley (1963) "Experimental and quasi-experimental designs for research on teaching", in: N. L. Gage (ed.) *Handbook of research on teaching*, Rand McNally, Chicago.
- Carvalho, S., and H. White (2004) "Theory-based evaluation: The case of social funds", *American Journal of Evaluation* 25(2), 141-160.
- Catley, A., J. Burns, D. Abebe and O. Suji (2008) *Participatory Impact Assessment: a guide for practitioners*, The Feinstein International Center, Tufts University, Medford.
- CGD (2006) *When will we ever learn? Improving lives through impact evaluation*, Report of the Evaluation Gap Working Group, Center for Global Development, Washington, DC.
- Chambers, R. (1983) *Rural development: putting the last first*, Wiley, New York.

- Chambers, R., (1995) "Paradigm shifts and the practice of participatory research and development", in: S. Wright and N. Nelson (eds.) **Power and participatory development: Theory and practice**, Intermediate Technology Publications, London.
- Chen, H.T. (1990) **Theory-driven evaluation**, Beverly Hills, Sage Publications.
- Cohen, J. and W. Easterly (eds.) (2009) **What works in development? Thinking big and thinking small**, Brookings Institution press, Washington D.C.
- Coleman, J.S. (1986) "Theory, social research and a theory of action", **American Journal of Sociology** 91(6), 1309-1335.
- Coleman, J.S. (1990) **Foundations of social theory**, Belknap Press, Cambridge.
- Cook, T.D. (2006) "Describing what is special about the role of experiments in contemporary educational research: Putting the 'Gold Standard' rhetoric into perspective", **Journal of Multidisciplinary Evaluation** 3(6), 1-7.
- Cook, T.D. and D.T. Campbell (1979) **Quasi-experimentation: Design and analysis for field settings**, Rand McNally, Chicago.
- Cousins, J.B. and E. Whitmore (1998) "Framing participatory evaluation", in: E. Whitmore (ed.) **Understanding and practicing participatory evaluation, New Directions for Evaluation** 80, Jossey-Bass, San Francisco.
- Davidson, E.J. (2006) "The RCTs-only doctrine: Brakes on the acquisition of knowledge?" **Journal of Multidisciplinary Evaluation** 3(6), ii-v.
- Deaton, A. (2009) "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development", **NBER Working Paper** 14690, Cambridge.
- De Leeuw, E.D., J.J. Hox and D.A. Dillman (eds.) (2008) **International handbook of survey methodology**, Lawrence Erlbaum Associates, London.
- Donaldson, S.I. and Lipsey, M.W. (2006) "Roles for theory in contemporary evaluation practice", in: I. Shaw, J.C. Greene and M.M. Mark (eds.) **The SAGE handbook of evaluation**, Sage Publications, Thousand Oaks.
- Duflo, E., and R. Hanna (2008) "Monitoring works: Getting teachers to come to school", **NBER Working Paper** 11880, Cambridge.
- Easterly, W. (2001) **The elusive quest for growth: Economists' adventures and misadventures in the tropics**, MIT Press, Cambridge.
- Ellis, F. (1988) **Peasant economics: Farm households and agrarian development**, Cambridge University Press, Cambridge.
- Elster, J. (2007) **Explaining social behavior - More nuts and bolts for the social sciences**, Cambridge University Press, Cambridge.
- Gill, G. (1993) **Ok the data is lousy, but it's all we've got (Being a critique of conventional methods)**, Gatekeeper Series 38, Sustainable Agriculture Program, International Institute for Environment and Development, London.
- Heckman, J. (1992) "Randomization and social program evaluation," **NBER Technical Working Paper** 107, Cambridge.

- Heckman, J., J. Smith and N. Clements (1997) "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts", *Review of Economic Studies* 64, 487-535.
- Hulme, D. (2000) "Impact assessment methodologies for microfinance: Theory, experience and better practice", *World Development* 28(1), 79-98.
- IFAD (2002) *Managing for impact in rural development: A practical guide for M&E*, IFAD, Rome.
- Jones, N., C. Walsh, H. Jones and C. Tincati (2008) *Improving impact evaluation coordination and uptake* - A scoping study commissioned by the DFID Evaluation Department on behalf of NONIE, Overseas Development Institute, London.
- Karlan, D. (2009) "Thoughts on randomised trials for evaluation of development: presentation to the Cairo evaluation clinic", *Journal of Development Effectiveness* 1(3), 237-242.
- Karlan, D. and J. Zinman (2009) "Expanding microenterprise credit access: Using randomized supply decisions to estimate impacts in Manila", *CEPR paper* 7396, London.
- Kusek, J and R.C. Rist (2004) *Ten steps to a results-based monitoring and evaluation system: A handbook for development practitioners*, World Bank, Washington D.C.
- Leeuw, F.L. (2003) "Reconstructing program theories: Methods available and problems to be solved", *American Journal of Evaluation* 24(1), 5-20.
- Leeuw, F.L. (2009) "On the contemporary history of experimental evaluations and its relevance for policy making", in: O. Rieper, F.L. Leeuw and T. Ling (eds.) *The evidence book: concepts, generation, and use of evidence*, Transaction Publishers, New Brunswick.
- Leeuw, F.L. and J. Vaessen (2009) *Impact evaluations and development – NONIE guidance on impact evaluation*, Network of Networks on Impact Evaluation, Washington D.C.
- Lipsey, M.W. (1993) "Theory as method: Small theories of treatments," in: L.B. Sechrest and A.G. Scott (eds.), *Understanding causes and generalizing about them, New Directions for Program Evaluation* 57, Jossey-Bass, San Francisco.
- Mark, M.M., G.T. Henry and G. Julnes (1999) "Toward an integrative framework for evaluation practice", *American Journal of Evaluation* 20(2), 177-198.
- McKenzie, D. and C. Woodruff (2008) "Experimental evidence on returns to capital and access to finance in Mexico", *World Bank Economic Review* 22(3), 457-482.
- Miguel, E. and M. Kremer (2004) "Worms: Identifying impacts on education and health in the presence of treatment externalities," *Econometrica* 72(1), 159-217.
- Mikkelsen, B. (2005) *Methods for development work and research*, Sage Publications, Thousand Oaks.
- Morgan, S.L. and C. Winship (2007) *Counterfactuals and causal inference – methods and principles for social research*, Cambridge University Press, Cambridge.
- Morra, L.G. and R.C. Rist (2009) *The road to results: designing and conducting effective development evaluations*, World Bank, Washington D.C.

- Oakley, A. (2000) *Experiments in knowing: Gender and method in the social sciences*, Polity Press, Cambridge.
- OECD-DAC (2002) *Glossary of key terms in evaluation and results based management*, OECD-DAC, Paris.
- OED (2005) *OED and impact evaluation: A discussion note*, Operations Evaluation Department, World Bank, Washington D.C.
- Olken, B. (2007) "Monitoring corruption: Evidence from a field experiment in Indonesia", *Journal of Political Economy* 115(2), 200-249.
- Pawson, R. (2006) *Evidence-based policy: A realist perspective*, Sage Publications, London.
- Pawson, R. (2009) "Middle range theory and program theory: from practice to provenance", in: J. Vaessen and F.L. Leeuw (eds.) *Mind the gap: perspectives on policy evaluation and the social sciences*, Transaction Publishers, New Brunswick.
- Pawson, R. and N. Tilley (1997) *Realistic Evaluation*, Sage Publications, Thousand Oaks.
- Pretty, J.N., I. Guijt, J. Thompson and I. Scoones (1995) *A trainers' guide to participatory learning and action*, IIED Participatory Methodology Series, IIED, London.
- Ravallion, M. (2008) "Evaluation in the practice of development", *Policy Research Working Paper* 4547, World Bank, Washington D.C.
- Ravallion (2009a) "Should the Randomistas rule?" *Economists' Voice*, February 2009.
- Ravallion, M. (2009b) "Evaluating three stylised interventions", *Journal of Development Effectiveness* 1(3), 227-236.
- Riggin, L.J.C. (1990) "Linking program theory and social science theory", in: L. Bickman (ed.) *Using program theory in evaluation, New Directions for Program Evaluation* 33, Jossey-Bass, San Francisco.
- Rodrik, D. (2008) "The new development economics: We shall experiment, but how shall we learn?" *MIMEO*, Harvard University, Cambridge.
- Rogers, E.M. (2003) *Diffusion of innovations*, New York, Free Press.
- Rogers, P. J. (2008) "Using programme theory for complex and complicated programs", *Evaluation* 14(1), 29-48.
- Rogers, P.J., Hacsı, T.A., Petrosino, A., and Huebner, T.A., (eds.) (2000) *Program theory in evaluation: Challenges and opportunities, New Directions for Evaluation* 87, Jossey-Bass, San Francisco.
- Rossi, P.H., Lipsey, M.W., and Freeman, H. E. (2004) *Evaluation: A systematic approach*, Sage Publications, Thousand Oaks.
- Salmen, L. and E. Kane (2006) *Bridging diversity: Participatory learning for responsive development*, World Bank, Washington D.C.
- Scriven, M. (2008) "Summative evaluation of RCT methodology: & an alternative approach to causal research", *Journal of Multidisciplinary Evaluation* 5(9), 11-24.
- Shadish, W. R., T.D. Cook and D.T. Campbell (2002) *Experimental and quasiexperimental designs for generalized causal inference*, Houghton Mifflin Company, Boston.
- Sherman, L.W., D.C. Gottfredson, D.L. MacKenzie, J. Eck, P. Reuter and S.D. Bushway (1997) *Preventing crime: what works, what doesn't, what's promising*, US Office of Justice Programs, Washington D.C.

- Skoufias, E. and B. McClafferty (2001) "Is PROGRESA working? Summary of the results of an evaluation by IFPRI", **FCND Discussion Paper 118**, IFPRI, Washington D.C.
- Tashakkori, A., and C. Teddlie (eds.) (2003) **Handbook of mixed methods in social and behavioral research**, Sage Publications, Thousand Oaks.
- The Lancet Editorial (2004) "The World Bank is finally embracing science", **The Lancet** 364(9436), 731-732.
- Vaessen, J. and D. Todd (2008) "Methodological challenges of evaluating the impact of the Global Environment Facility's biodiversity program", **Evaluation and program planning**, 31(3), 231-240.
- Vaessen, J., F. Leeuw, S. Bonilla, R. Lukach and J. Bastiaensen (2009) "Protocol for synthetic review of the impact of microcredit", **Journal of development effectiveness**, 1(3), 285-294.
- Van der Knaap, L.M., F.L. Leeuw, S. Bogaerts and L.T.J. Nijssen (2008) "Combining Campbell standards and the realist evaluation approach – the best of two worlds?", **American Journal of Evaluation** 29(1), 48-57.
- Weiss, C.H. (2000) "Which links in which theories shall we evaluate?" in: P.J. Rogers, T.A. Hacsí, A. Petrosino and T.A. Huebner (eds.) **Program theory in evaluation: Challenges and opportunities, New Directions for Evaluation** 87, Jossey-Bass, San Francisco.
- White, H. (2009) "Some reflection on current debates in impact evaluation", **Working Paper 1**, International Initiative for Impact Evaluation, New Delhi.
- Woolcock, M. (2009) "Toward a plurality of methods in project evaluation: a contextualized approach to understanding impact trajectories and efficacy", **Journal of Development Effectiveness** 1(1), 1-14.
- Worrall, J. (2007) "Why there's no cause to randomize", **The British Journal for the Philosophy of Science** 58(3), 451-488.



University
of Antwerp



INSTITUTE OF DEVELOPMENT
POLICY AND MANAGEMENT